R E P O R T R E S U M E S

ED 012 956 EA 000 548

EMPIRICAL TAXONOMIES OF FOUR-YEAR COLLEGES AND UNIVERSITIES.

BY- CREAGER, JOHN A.

PUB DATE FEB 67

EDRS PRICE MF-$0.25 HC-$0.52 13P.

DESCRIPTORS- *HIGHER EDUCATION, MODELS, *CLASSIFICATION, COMPUTER ORIENTED PROGRAMS, ORGANIZATION, METHODOLOGY, *ENVIRONMENTAL INFLUENCES, *INPUT OUTPUT, CHARTS, *INSTITUTIONS, NEW YORK,

A MODEL OF HIERARCHICAL GROUPING IS APPLIED TO 24 INSTITUTIONS OF HIGHER EDUCATION. THE GROUPING IS A FUNCTION OF DIFFERENCES AMONG THE INSTITUTIONS' CHARACTERISTICS. THE THREE HIERARCHICAL GROUPINGS SHOWN ARE BASED ON (1) 10 USOE CATEGORIES OF INSTITUTIONAL CHARACTERISTICS, (2) 14 INPUT AND ORIENTATION VARIABLES, AND (3) 36 COLLEGE ENVIRONMENT VARIABLES. THE FINDINGS SHOW THAT AN INCREASE IN EITHER THE NUMBER OF INSTITUTIONS OR THE NUMBER OF VARIABLES MAY RESULT IN MORE GROUPS BEING DEFINED AT THE CRITERION LEVEL. FURTHER METHODOLOGICAL ISSUES RAISED INCLUDE THE EXTENDED GROUPING OF 245 INSTITUTIONS. THE SIZE OF THE STUDY POPULATION IS LIMITED ONLY BY COMPUTER CAPACITY AND AVAILABLE DATA. THIS PAPER WAS PRESENTED AT THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION ANNUAL MEETING (NEW YORK, FEBRUARY 16-18, 1967). (HW)

ED012956

Empirical Taxonomies of Four-year Colleges and Universities[1]

John A. Creager
American Council on Education

In studies of higher education it is often desirable to classify institutions. This paper describes the application of hierarchical grouping and associated computer algorithms to matrices of distances among institutions in spaces defined by various sets of variables. The studies being reported are based on 24 institutions deliberately selected to illustrate the features of the model. I shall also discuss certain methodological issues regarding the choice of distance measures, the metrics of the variables, and the problem of intercorrelation among the variables.

Hierarchical grouping starts with a paired-comparisons matrix of some measure of the similarities or differences among a set of objects. In our studies the objects are institutions and the input matrix elements are measures of differences among institutions. For a given pair of objects this measure is defined by the sum of the squared differences between the two objects along each axis, or variable, in a Euclidean space.

From the matrix of distances the hierarchical grouping model classifies the institutions into the pattern familar in biology, i.e., into species, genera, families, orders and kingdoms. The full hierarchical grouping elaborates this pattern into a total of n-1 levels of the hierarchy. One may be interested in the entire hierarchical pattern, or in the groups formed at a given level of the hierarchy. At the first level the two objects most similar in terms of the input measures are clustered and all others are single-membered groups or isolates;

---

[1]Presented at the meeting of the American Educational Research Association, New York, February 16-18, 1967.

EA 000 548

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

at the final level the entire set of objects is in a single group. The intermediate levels exhibit a number of groups of varying size.

At each stage, a cluster is formed, the cluster centroid is computed, and comparisons of this cluster are made with each other cluster including isolates. The grouping algorithm minimizes an objective function, which measures the loss of information implied by the grouping at that stage. Ward and Hook (8) have defined an objective function for distance matrices in terms of within-group variation about group centroids.

Figure 1 displays the results of grouping the 24 institutions in a space defined by 10 institutional characteristics used by the USOE.[2] The grouping stages are indicated by numbers in the brackets. One of three things may occur at any stage of the grouping. First, two isolates may be brought together to form the nucleus of a new cluster. This is more likely to occur in the early stages of grouping. For example the clustering of two isolates is illustrated by Colorado State College and Nebraska State Teachers College clustering at stage one, and by Yale and Dartmouth coming together at stage four. The second alternative at any given stage is that two previously formed clusters may come together to form a more heterogeneous cluster. This may occur anywhere in the process except at stage one, but is most common in the later stages of the grouping, for example in stages 18 through 23. Third, an isolate may be assigned to a previously formed cluster, as in steps 2 and 3 where Lock Haven State College and Southwest Texas State College are added, respectively, to the cluster defined at step one.

For most purposes we are less interested in the entire hierarchical structure than we are in the number and makeup of the groups at the level of maximum reduction of the number of objects, with minimum loss of information. Ward (7)

---

[2] Men's, women's, universities, liberal arts, teachers, technical, Protestant, Catholic, private nonsectarian, and predominantly Negro.

has suggested that we consider the level where there is a sharp break in either the objective function or in the change in the objective function. Employing this criterion we accepted the grouping at level 17, shown by the vertical line drawn down through the cluster brackets. At this level we find no isolates remaining and seven groups defined by the Roman numerals to the left of the figure. Clearly such groups are meaningful with respect to the input variables on which the grouping was based.

The result of grouping the same 24 institutions on distances in a different description space is presented in Figure 2. Here we used 14 variables: five freshman input factors; six Environmental Assessment Technique Orientations; size; selectivity; and affluence. The orientations are measures of curricular emphases; affluence is measured by per student expenditures for general and educational purposes (4). Many of these variables are related to the 10 variables used in the previous analysis. Hence, it is not surprising to find the resulting grouping of 24 institutions similar to that just examined. Cutting the hierarchy at step 16, we find eight clusters instead of seven. Seven of the eight clusters are essentially the same as before. The university cluster found in the first analysis has been split into the large midwestern state universities and a new group consisting of the A. & M. type of school. Illinois Institute of Technology has moved into this group from the previous technical group, being more similar to the A. & M. group in terms of Intellectualism and Status Orientation. The College of Our Lady of the Lake has clustered with the teachers colleges due to the high proportion of students in teacher preparation. The other two Catholic colleges have a more liberal arts orientation (5). Southwest Texas State College, which was in the teacher's group in the first analysis, appears here with the Protestant liberal arts group, since this institution is more similar to this group on Size, Masculinity, and Status Orientation.

In the third analysis of the same 24 institutions, we used 36 environmental variables to define the description space. These variables were identified by Astin (2) in a recent study of college environments. Twenty-eight of the variables represent factors based on potential stimuli capable of changing the students' sensory input.[3] The other eight variables were developed from items similar to those in the College and University Environmental Scales (6). The results of this analysis are presented in Figure 3. The objective function criterion was met at level 18 with only six groups, which were felt to be too heterogeneous. After examining the grouping, it was decided to cut at level 16 where there are eight groups and no isolates. Since there is some correlation between the sets of variables defining the description spaces used in the three analyses (3), the grouping shown in Figure 3 is similar to those observed in the first two analyses.

Our experience with several analyses indicates that an increase in either the number of institutions or the number of variables may result in defining more groups at the criterion level of the hierarchy, and that the choice of a cutting point may become more difficult. However, an increase in the number of input variables may not increase the number of groups where there is appreciable correlation among the variables.

At this point I would like to discuss the methodological issues raised by the problems of differences in metric and of the intercorrelations among the variables. The Pythagorean distance between any two points in the description space is a function of the metrics of the input variables. Where these metrics are arbitrary and where each variable is considered equally important for classification, division of the input scores by their standard deviations is recommended. Failure to do so has the effect of placing greater weight on those variables with large variances and underweighting those with small variances. The effect of this on the distance

---

[3]These include measures of the peer environment, the classroom environment, the physical environment, and the administrative environment. The items on which the measures are based can be verified by independent observation.

measures and on the grouping may be seen in Figure 4. In this figure, four hypothetical institutions are located in two dimensions. Points one and two, and points three and four, form the two groups, A and B. After equating for metric, the relative locations of the points have shifted, as shown by the arrows, to the new positions indicated by the primes. Points 2', 3', and 4' now form cluster C and leave point 1' as an isolate. In the three analyses reported here, the computer algorithm equated the metrics of the input variables. If any rational basis exists for weighting input variables in computing distances, the original metrics can be used. I judge that this will not usually be the case, although I shall shortly discuss an exception.

The general effect of correlations among the variables is to weight the common factors in the computation of the distances. In the extreme case, were we to include the same variable twice, the distance measure for each pair of institutions would be doubly weighted on this dimension. The effect on subsequent grouping depends on the pattern of projections for the whole set of institutions on this dimension, and upon the relative importance of this dimension to the others in discriminating the institutions. A case can be made, however, for common factor weighting where the input variables are well chosen and do not involve artifactual linear dependencies.

The general answer to the treatment of correlated variables is to use the Mahalanobis distance function. This corrects for both correlation and metric. However, it requires the inversion of the dispersion matrix, which may be singular. Where the inverse does exist, the same hierarchical grouping can be obtained from the Pythagorean distances in the space defined by the principal components of the correlations among the original variables. For this equivalence between Mahalanobis and principal components to hold, it is necessary to equate the metrics of the institution's scores on these components. This, however, may be criticized, and by implication, the Mahalanobis distance basis for grouping also criticized, because one would not ordinarily wish to give equal weight to all components, either for

computing the distances or for the subsequent grouping. To do so is to give equal weight to those components which are largely error variance. It is therefore thought preferable to leave the scores on the principal components in the metrics resulting from multiplying the original z scores by the eigenvectors. Thus, each component would contribute to grouping in proportion to the amount of variance it contributes to the original description space. In view of the reliability of component scores, as measured by the alpha coefficient, a case can be made for using only those components with latent roots greater than one. These issues, involving the factorial content and degree of redundancy in the input variables, are currently being investigated.

Finally, there is one other methodological issue, as it relates to the taxonomy of colleges and universities. We are really interested in grouping 245 institutions for which we have extensive data. Although there is nothing in the logic of the model which precludes grouping such large sets, the amount of computer storage and the running time both increase geometrically with arithmetic increase in the number of institutions. One solution to this problem involves forming clusters on one or more samples of the larger set and combining results across samples or with hierarchical grouping of clusters. The results would not necessarily be the same as those obtained from grouping on the total set, but the number and makeup of the final groups may be sufficiently acceptable for practical purposes and well worth the saving of computer costs.

With some of these methodological issues under control, our future effort will be focused on the grouping of the larger sets. Our studies to date indicate that classification of a small, highly structured set of institutions is rather stable across different input variables and treatments. Some doubt remains about this stability in larger, and less well structured sets.

Hierarchical Grouping of 24 Institutions on Ten USOE Categories

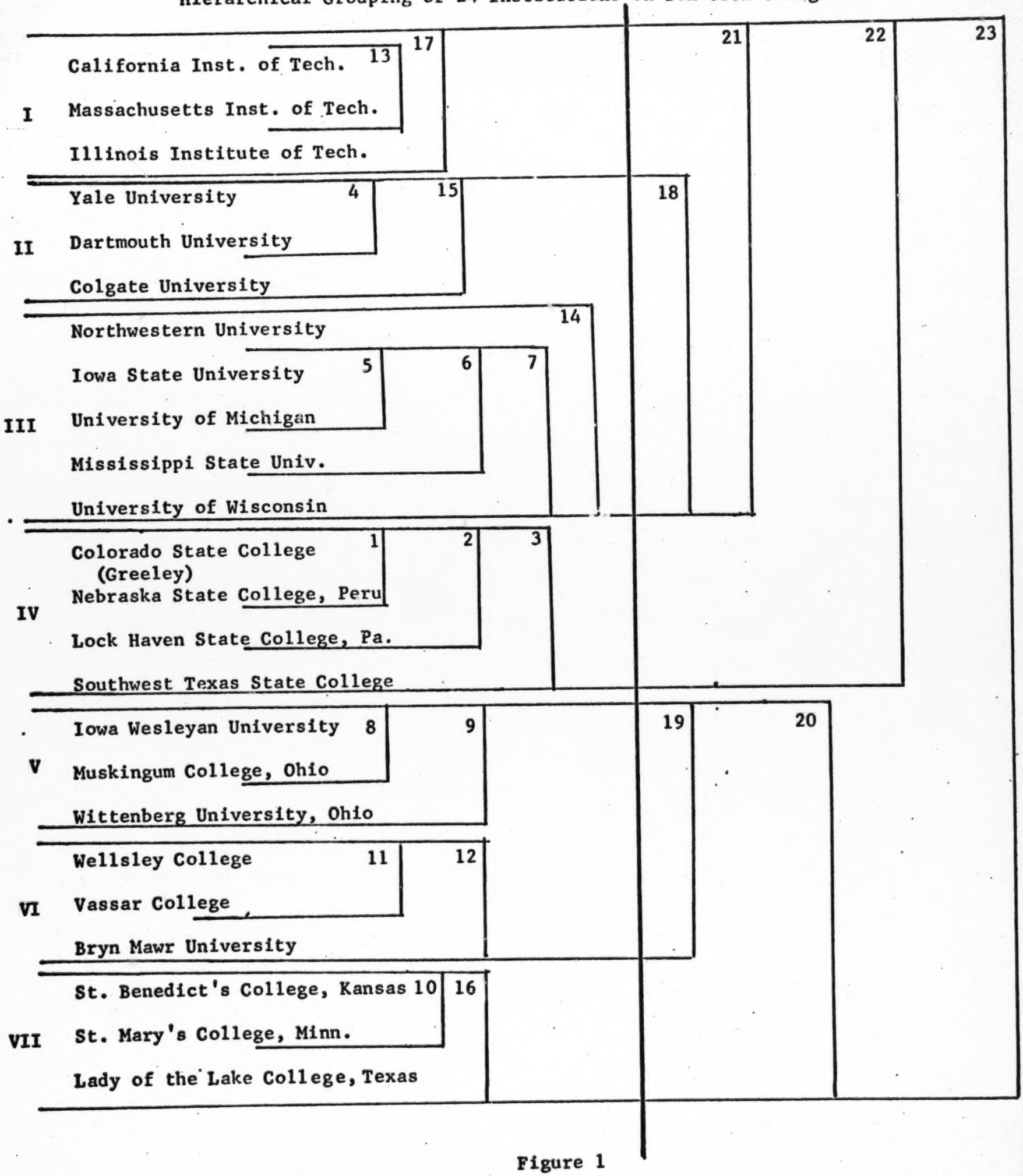

Figure 1

Hierarchical Grouping of 24 Institutions on 14 Input and Orientation Variables

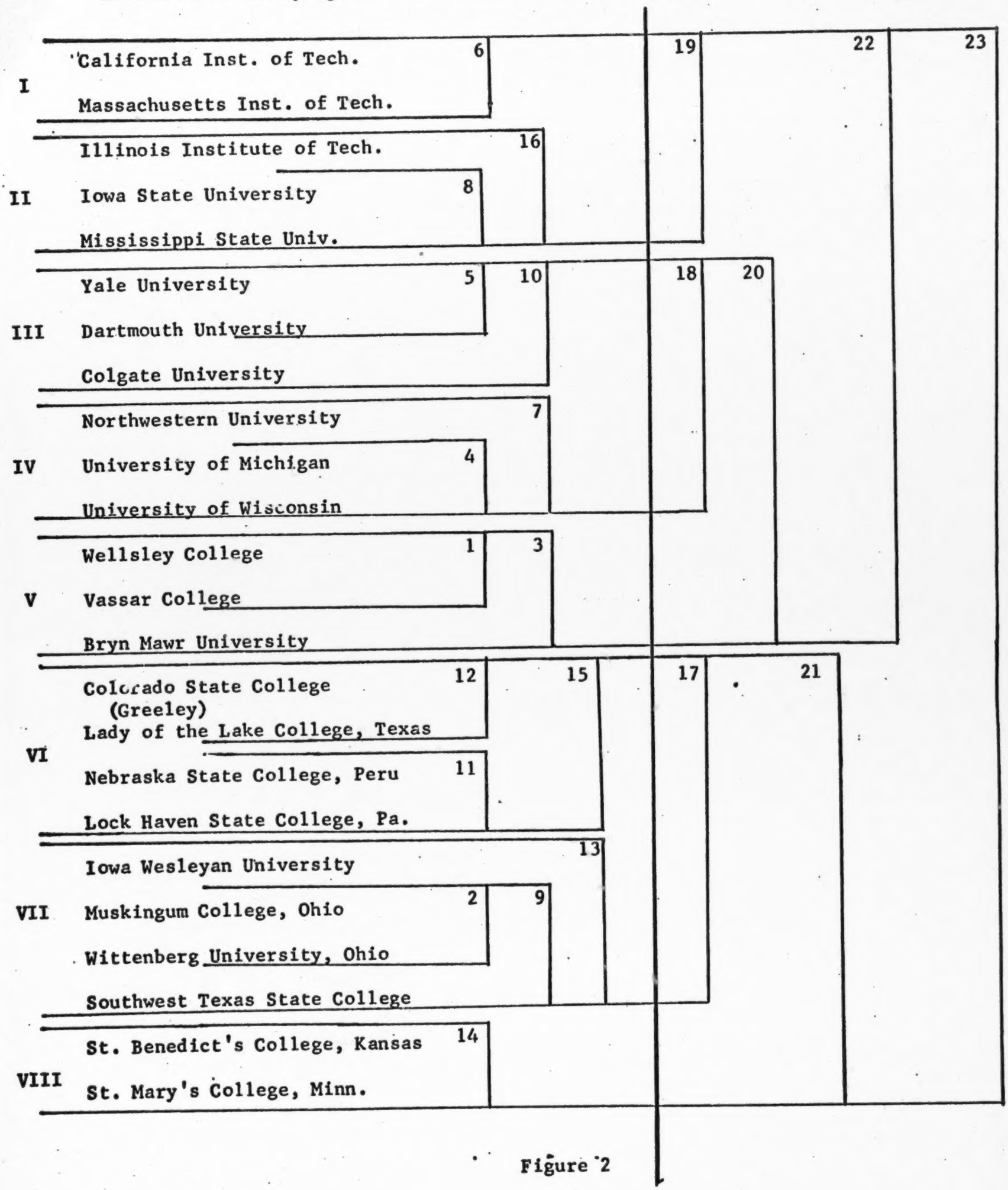

Figure 2

Hierarchical Grouping of 24 Institutions on 36 College Environment Variables

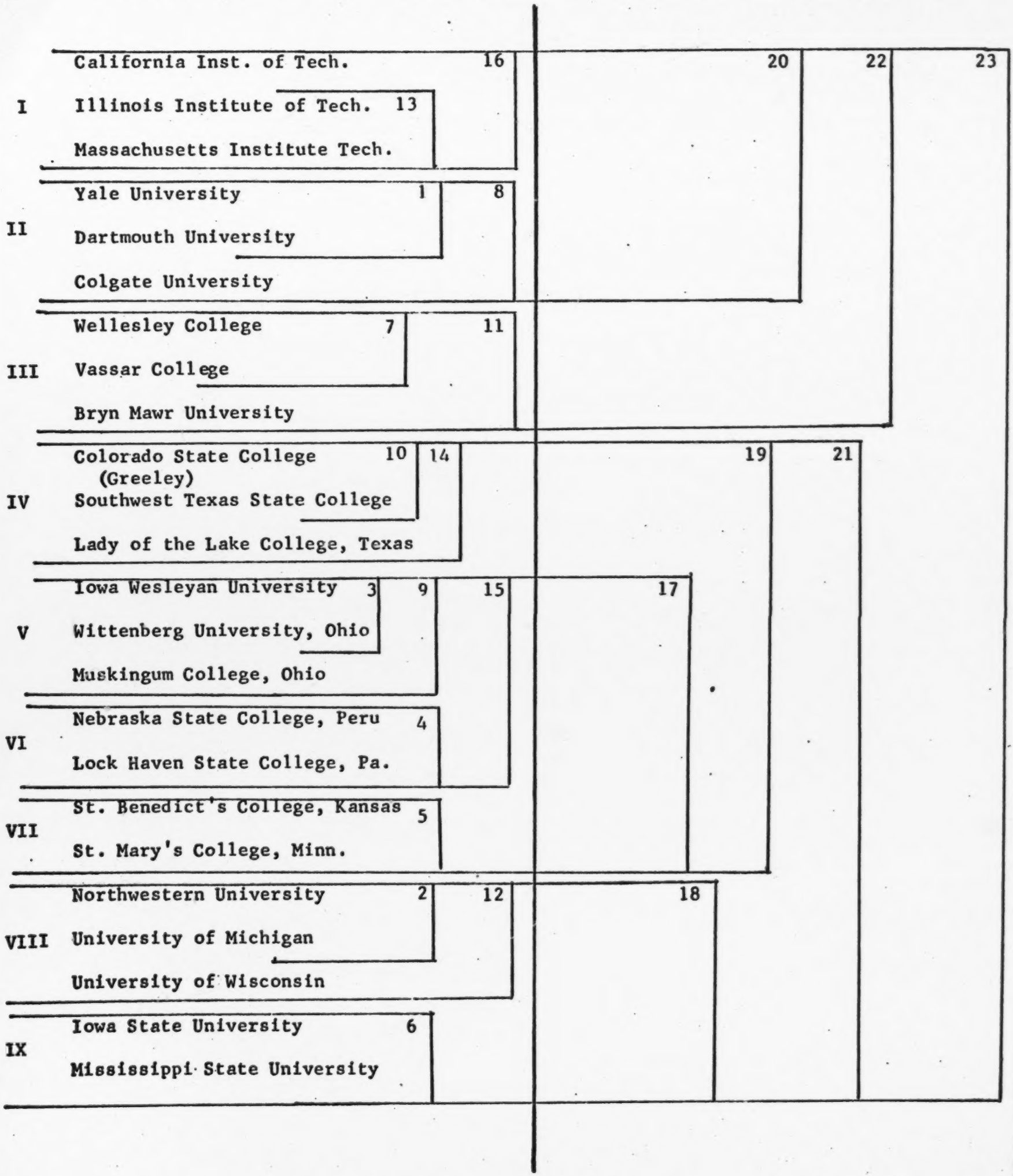

Figure 3

Effect of Metric on Interpoint Distance

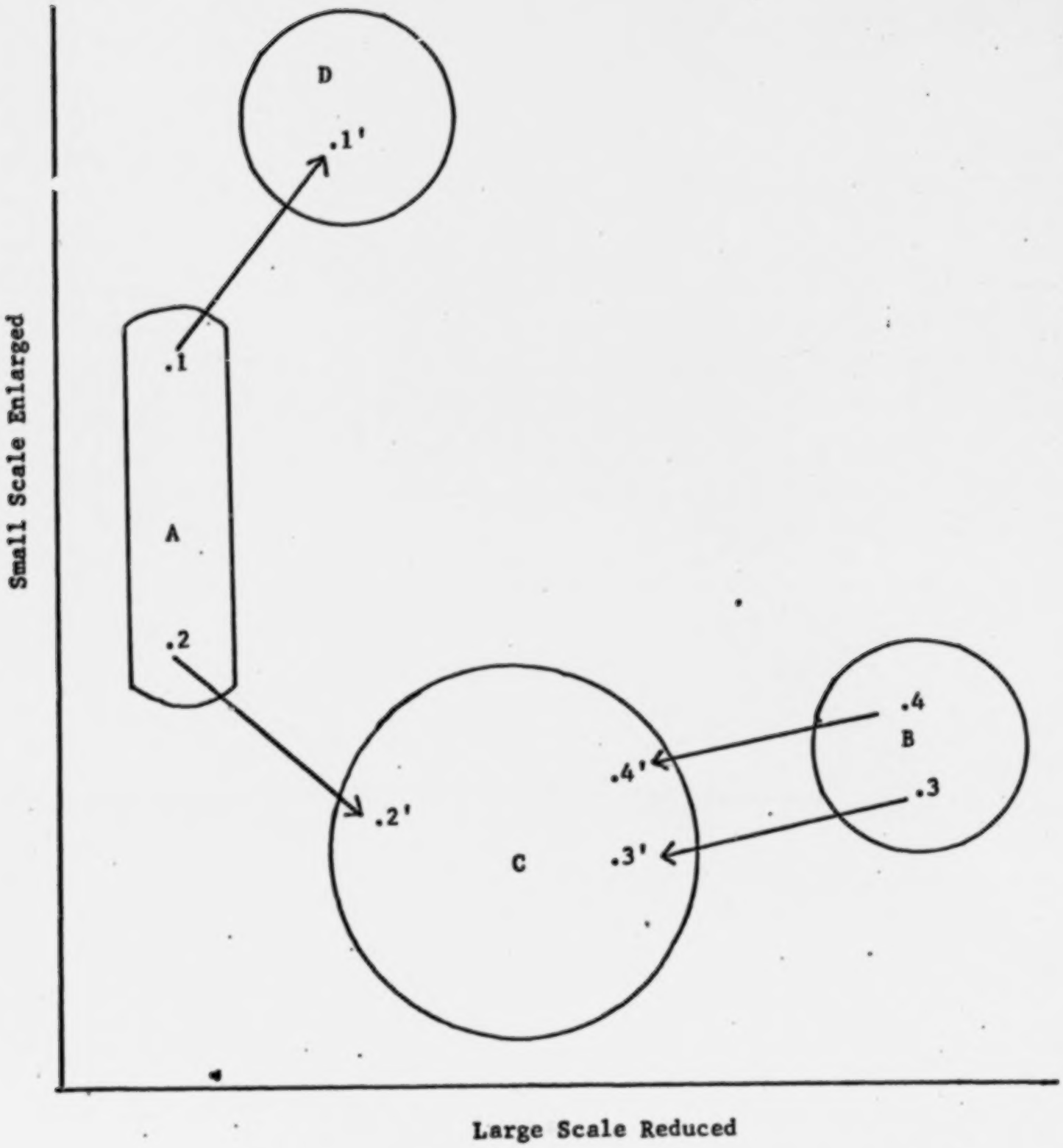

Figure 4

References

1. Astin, A. W. Who Goes Where to College? Chicago: Science Research Associates, Inc., 1965.

2. Astin, A. W. The College Environment. Unpublished manuscript.

3. Astin, A. W., and Creager, J.A. "Alternative Methods of Describing Charateristics of Colleges and Universities," Unpublished manuscript.

4. Cartter, A. M.(ed.) American Universities and Colleges. Washington, D. C.: American Council on Education, 1964, 1278 pp.

5. Office of Education. Higher Education, Part 3, Washington, D. C.: U.S. Government Printing Office, 1966, 224 pp.

6. Pace, C. R. "The Influence of Academic and Student Subcultures in College and University Environments," Washington: U.S. Office of Education Cooperative Research Project No. 1083, 1964, 225 pp.

7. Ward, J. H., Jr."Hierarchical Grouping to Maximize Payoff." Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Division, Air Research and Development Command, USAF, March 1961. (Technical Note WADD-TN-61-29)

8. Ward, J. H., Jr. and Hook, M. E. "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles," Educational and Psychological Measurement Journal, Vol. XXIII, No. 1, 1963, pp 69-81.